Review of *Sparse inverse covariance matrix with the graphical lasso*

Cameron Davidson-Pilon

Nov. 1st, 2012

Abstract

We review the Graphical Lasso algorithm by Hastie *et al.* and use it to explore the complex relationships in the equity markets.

1 Introduction

We have seen how important estimation of parameters can be when using them as inputs for an optimization program, and how optimizers can be error maximizers. The old statistics adage *garbage in, garbage out* can be aptly modified to *garbage in, really bad garbage out*. Furthermore, often we want to be able to add more structure to our estimates. This added structure can come from domain knowledge of the subject or prior beliefs about the estimates. The authors of [1] introduce a novel algorithm to estimate the inverse of an additionally structured covariance matrix. The algorithm, called Glasso or graphical lasso, is used to impose penalty conditions on the maximum likelihood estimator of the inverse covariance matrix.

1.1 Interpretation of the inverse correlation matrix

The inverse correlation [1] matrix, Θ, has a specific mathematical interpretation. The element $\Theta_{i,j}$ is proportional to the *partial correlation* between variables x_i and x_j. The partial-correlation between variables X and Y, given Z, denoted $\rho_{X,Y|Z}$, is defined as the correlation between the residuals R_x and R_y after least-squares regressing X on Z and Y on Z, respectively. One can think of this as a measure of how correlated two vectors are given the influence of a set of other variables has been considered. An important result concerns if the vectors X and Y are normally distributed, in which case a partial correlation of 0 implies conditional independence.

[1] All this extends to the inverse covariance matrix, but for simplicity's sake we consider the correlation matrix

Figure 1: Left: The partial correlation, after considering the group, is about equal to the correlation. Center: The partial correlation is greater than the correlation. The partial correlation reveals a spurious correlation. Figures from [6].

The relationship between linear regression residuals and the inverse covariance matrix is at first thought surprising, but the relationship is more clear once one recalls the solution to the linear least-squares problem is

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= \Theta X^T Y\end{aligned}$$

1.2 Motivating example

Consider the absolute daily stock returns of PetSmart (PETM) and Rowan Companies (RDC). [2] The correlation between the returns over the period Jan 2009- Oct 2012 is 0.455 (the t-test of non-correlation fails with p-value = 9.4e-9). If one has information that PETM would increase by 10% tomorrow with certainty, one would not feel comfortable investing in a gas company too, even though they have a significant positive correlation. Heuristically, the hesitancy to invest is that there is no economic reason for a relationship to exist between a pet-supply store and an oil drilling company. Mathematically, the hesitancy to invest is that correlation hides causation. In this example, neither PETM or RDC causes the other to move but a third variable, the S&P 500, is actually the main factor in the high correlation. Figure 2 shows a toy example of this.

What would be desirable is to estimate a covariance matrix, or more appropriately an inverse correlation matrix, that can detect and remove elements that have reflect a spurious correlation.

[2] PetSmart is a retail distributor of pet supplies and pet services. Rowan Companies provides contract oil well drilling services and rigs.

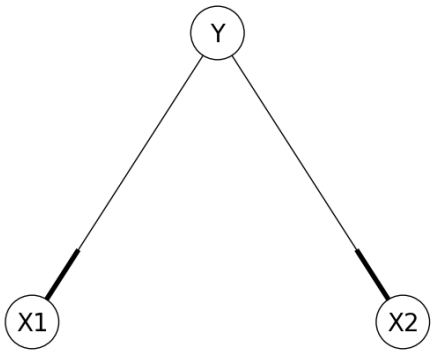

Figure 2: As Y is the factor that controls X_1 and X_2, there likely exists a correlation between X_1 and X_2, though the two are independent when conditioned on $Y \Leftrightarrow$ the partial correlation is 0 if the variables are normally distributed.

2 Lasso-type problems

Unfortunately, elements of the estimated covariance matrix are never zero and thus all elements of the partial-correlation matrix are never zero. But often, like in the financial example above, one has reason to believe that there exist null partial-correlations. One way to induce null partial-correlations is to penalize the maximum likelihood estimate of the inverse covariance matrix using an $L1$ penalty function:

$$\max_{\Theta} \quad \log\det\Theta - \mathrm{tr}(S\Theta) - \alpha||\Theta||_1 \tag{1}$$

over non-negative definite Θ, where $||\Theta||_1$ is the sum of the absolute value of elements of Θ and S is the empircal covariance matrix. The specific use of the $L1$ penalty here has the interesting effect of being able to set elements to 0, creating a sparse matrix. Contrast this with an $L2$ penalty which will shrink elements but never reduce them to zero. The sparsity of the matrix is desirable not only for computational reasons but also for parsimonious reasons: by removing elements, a simpler theory of relationships results.

The use of the $L1$ penalty in linear regression is known as a *Lasso problem* [3]. The basic Lasso problem looks like

$$\min_{\beta} \quad ||Y - X\beta||^2 + \alpha||\beta||_1$$

The solution is a sparse vector resulting from the addition of the $L1$-norm penalty term. See figure 3, 4.

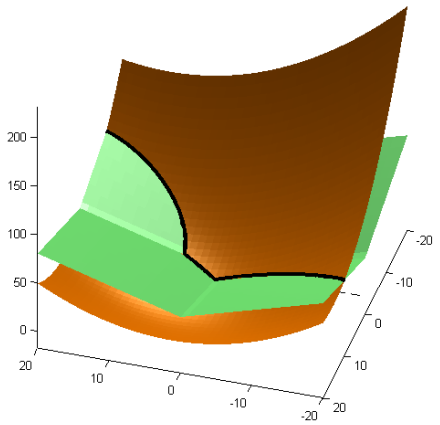

Figure 3: In a two-dimensional Lasso regression problem, the minimization occurs one the line of intersection of the least-squares function and the $L1$ function.

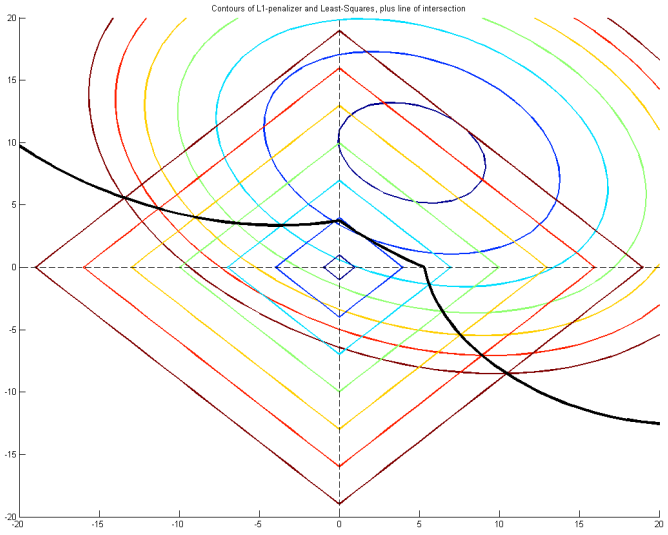

Figure 4: The contour plot of the surfaces in 3. Figures from [2]

2.1 Glasso algorithm

The authors of [1] developed a fast algorithm, referred to as Glasso, to solve the optimization problem 1. Let W be the estimate of Θ^{-1}, and S is the standard correlation matrix estimate. The following notation will be useful in what follows

$$W = \begin{pmatrix} W_{1,1} & w_{1,2} \\ w_{1,2}^T & w_{2,2} \end{pmatrix}$$

$$S = \begin{pmatrix} S_{1,1} & s_{1,2} \\ s_{1,2}^T & s_{2,2} \end{pmatrix}$$

The authors show that the solution for $w_{1,2}$ is equal to

$$\min_{\beta} \quad \frac{1}{2}||W_{1,1}^{\frac{1}{2}}\beta - b||^2 + \alpha||\beta||_1 \tag{2}$$

where $b = W_{1,1}^{-\frac{1}{2}} s_{1,2}$. Then, if β solves the above, $w_{1,2} = W_{1,1}\beta$. This presupposes we know $W_{1,1}$, but we only have our current estimate. Thus value $w_{1,2}$ is not the final solution, so we must iterate this procedure. In [1], the authors propose the following algorithm.

1. Let S be the p by p covariance matrix. Start with $W = S + \rho I$. The diagonal of W remains unchanged in what follows.

2. For each j = 1,2,...p, 1, 2, ..., p, ... solve the lasso problem 2. This gives a $p-1$ solution $\hat{\beta}$. Fill in the corresponding row and column of W using $w_{1,2} = W_{1,1}\hat{\beta}$.

3. Repeat until convergence.

4. Compute $\hat{\Theta} = W^{-1}$

The authors of [1] also propose to stop iterations when the average absolute change in W is less than $t \times \text{ave}(S^{-\text{diag}})$ where $S^{-\text{diag}}$ are the off diagonal elements of the empircal covariance matrix and t is a threshold, set by default to 0.001.

Notice that if we set $\alpha = 0$, then step 1 is simply $W = S$ and if we proceed with the above algorithm, then one sweep through the predictors computes S^{-1} using standard linear regression.

2.2 Recent improvements on Glasso algorithm

Since the introduction of the Glasso algorithm, improvements by the the initial authors and others have been discovered. These improvements are subjects to their own reviews, but can be summarized as follows:

1. improving stability by shrinking the initial covariance estimate.
2. improved stopping condition by examining the duality gap [4].
3. improved performance when the estimated solution is a block-diagonal matrix [5].

Our implementation uses the first two advances.

3 Applications

3.1 Finance

The authors of [1] use the Glasso algorithm on micro-array data. We instead choose to apply the algorithm to time series data, in particular the recent history of the equity market.

Our original problem was trying to unknot the relationships between stock price returns by shrinking partial-correlations to zero. One can think of a partial-correlation matrix as a large network with connections, or edges, when the partial correlation is non-zero. By "pruning" the partial-correlation matrix, one can estimate variable clusters and the network. This enables a practitioner or manager to answer questions like

- What the are the clusters of stock return?
- What are the exposures to my exposures?
- What hidden relationships exist in the network?

For this study, we examine different price returns from a subset of the equity market between the period January 3rd, 2010 and October 28th, 2012. The following companies' absolute stock returns are used:

MS	TOT	F	TM	MTU	TWX	CVX	MAR	MMM	HMC
SNE	CAJ	BAC	K	PFE	XRX	AIG	PEP	KO	PG
MCD	WMT	JPM	C	WFC	GE	T	VZ	IBM	MSFT
GOOG	AAPL	RIMM	^GSPC	CSCO	YHOO	ORCL	SNDK	DELL	NVDA
EBAY	AMD	S	INTC	VXX					

We standardize all the time series to have zero mean and unit variance. We perform cross-validation to find an appropriate α value. For all experiments, $\alpha = 0.25$ unless otherwise stated.

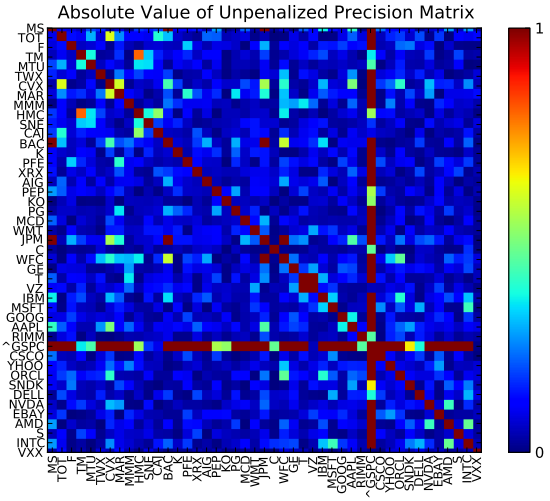

Figure 5: The unpenalized precision matrix of 45 equities.

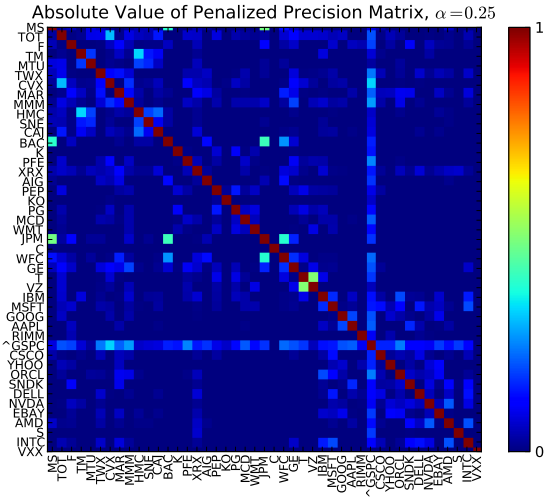

Figure 6: The penalized precision matrix of 45 equities after applying Glasso with $\alpha = 0.25$. Note how many entries are zero, and whatever is nonzero shows significant relationships between the stocks. It is clear that a major driver of most other stocks is the S&P 500 (represented by ^GSPC above). Similarly, note how the influence of WFC diminishes greatly.

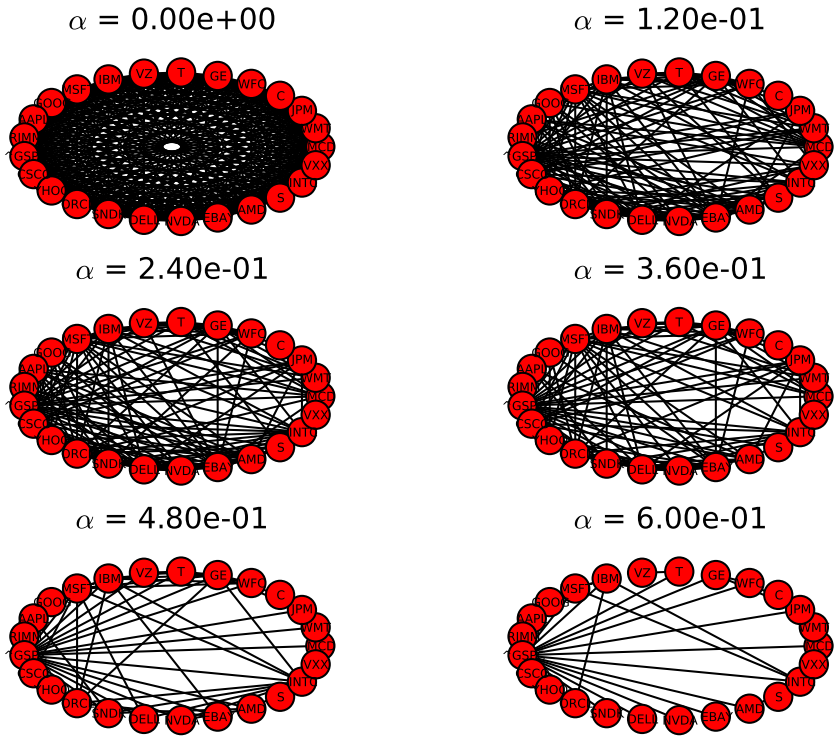

Figure 7: An edge between two nodes represents a non-zero partial-correlation. If $\alpha = 0$, the network starts as a fully connected network. As α increases, spurious relationships are pruned. The final α value shows what strong dependence the S&P 500 has.

By employing the Fruchterman-Reingold force-directed algorithm to our network, we can recover a two dimensional representation of the time series' relationships. The Fruchterman-Reingold force-directed algorithm works by treating the edges as springs with forces proportional to the partial-correlation, and iterating the springs' dynamics. Figures 8 and 9 demonstrate this.

What sort of interesting relationships can we recover from the exposed network structure shown in figure 9? It is clear that industries group together, but what's more interesting are the companies that "bridge" the industries. For example, ACE in an insurance company that is visually stuck between the financial cluster and the energy sector. Similarly, TOT (Total

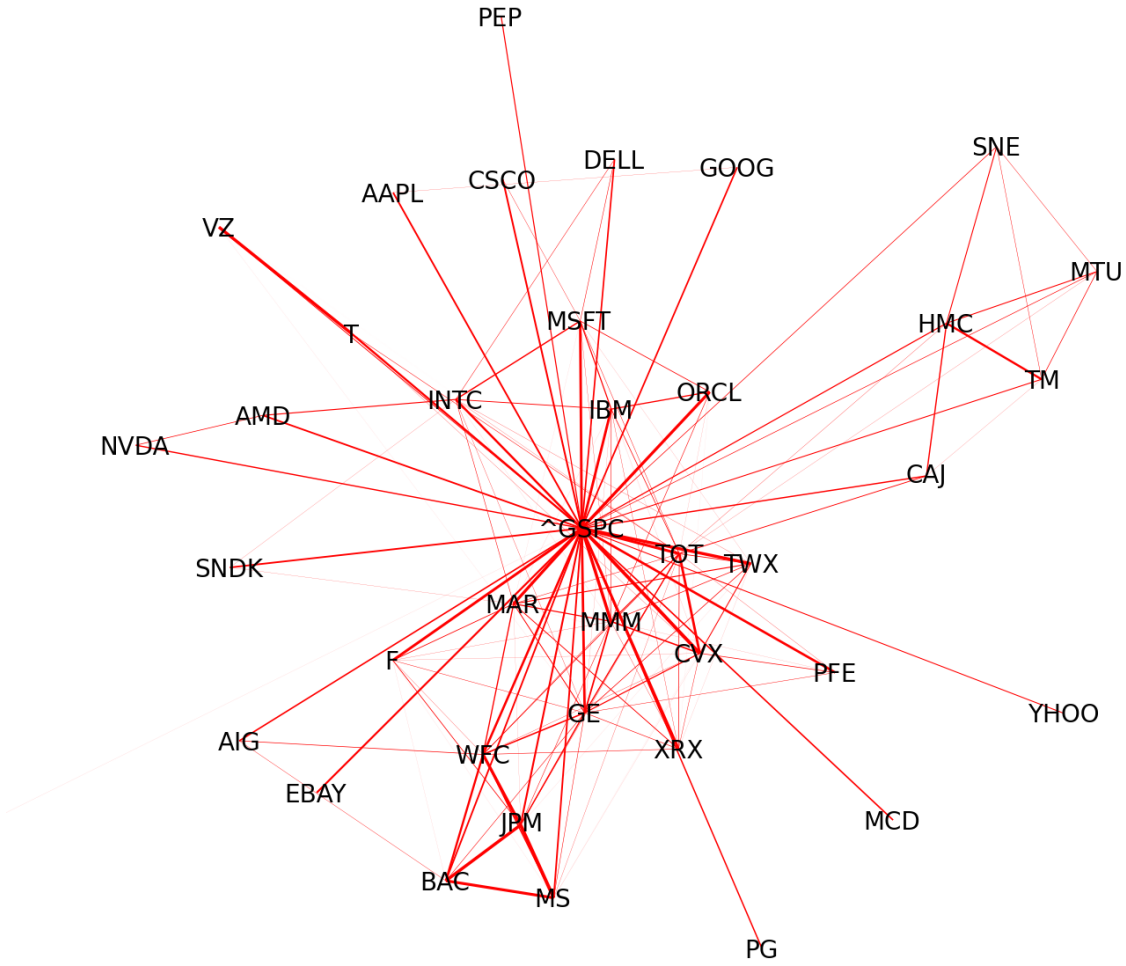

Figure 8: By employing the Fruchterman-Reingold force-directed algorithm to our network, we can recover a 2D representation of the time series' relationships. The weight of an edge is proportional to the partial-correlation. The north-east corner contains the Japanese auto and tech industry; the most northern nodes are all software and hardware tech; there is a tight connection between JPM, MS, BAC and WFC in the south. The Fruchterman-Reingold force-directed algorithm works by treating the edges as springs with forces proportional to the partial-correlation, and iterating the dynamics. $\alpha = 0.52$

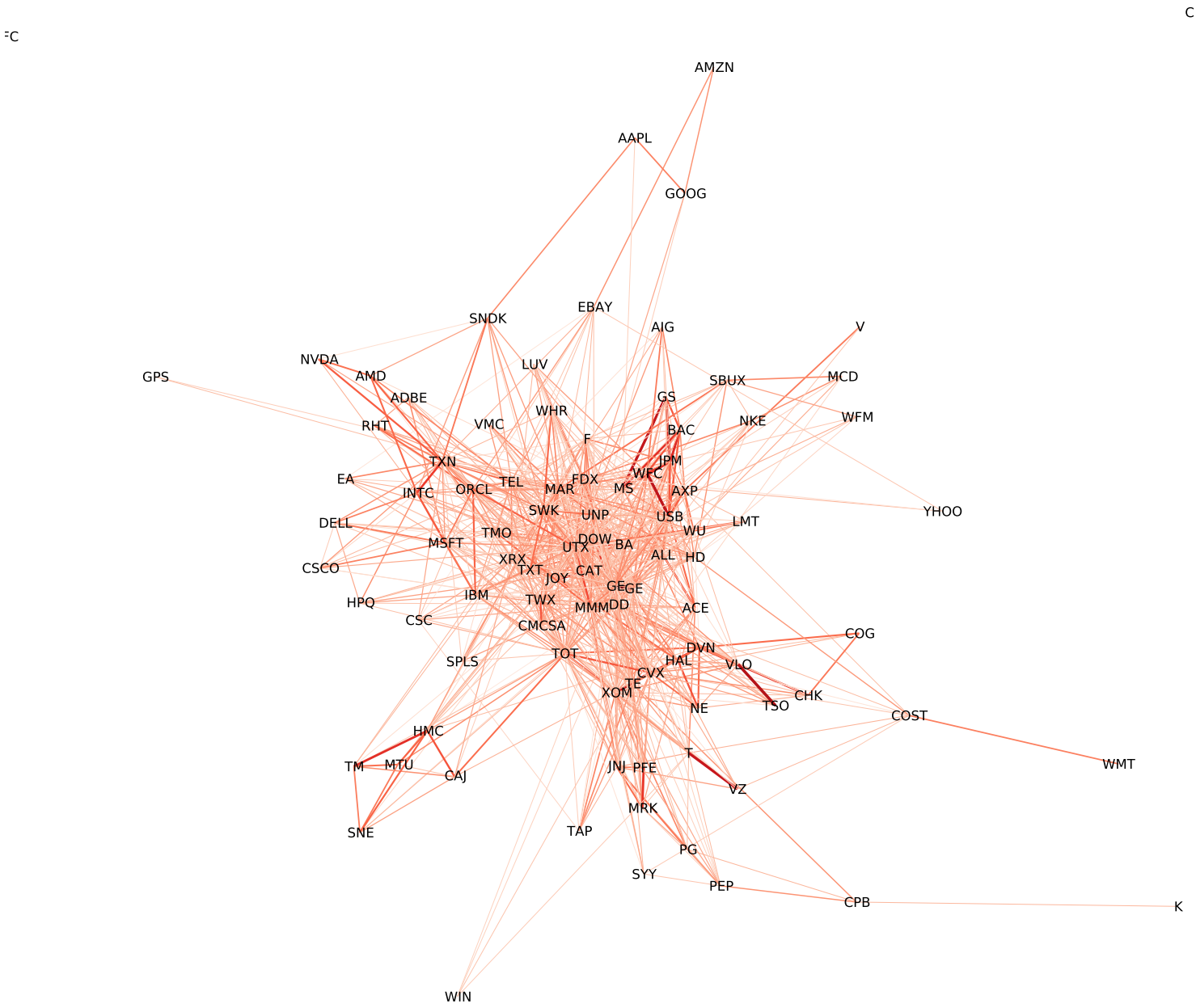

Figure 9: The result of applying the Fruchterman-Reingold force-directed algorithm to a larger set of timeseries. The S&P500 has been removed for clarity. A darker, bolder edge indicates a larger partial-correlation. One can pick out groups easily. The Financial District is back (GS, BAC, WFC, and others), a health-science group (JNJ, PFE, MRK), the Japanese Penisula (TM, HMC etc.), large web-service companies (GOOG, AAPL, AMZN), computer hardware (TXN, AMD, NVDA, etc.), next to computer assemblers (DELL, CSCO, HPC, MSFT, etc.), and the energy giants (TE, TOT, COG, TSO, NE, etc.).

Energy, based in Calgary, Alberta) seems to be one of the few companies that is connected to the Japanese peninsula.

This technique does not appear to work well for smaller companies (relative to their proportion in the S&P 500). Companies like WIN, GPS, TAP and MAR appear to be misplaced in the diagram. Though there are some larger companies that look misplaced too. It is interesting that F (Ford Motor Company) is very closely positioned to the Financial district. One can speculate why.

It is easy to see what the *exposures to exposures* are. For example, a manager investing in Bank of America (BAC) or Morgan Stanely (MS) would probably do well to hedge his or her exposures with another member of the financial district. Or if investing in TOT (Total Energy), it would be wise to consider the economic health of Japanese companies.

4 Conclusion

We have explored and implemented a slightly more stable Glasso algorithm than the algorithm outlined in [1]. While the initial motivation for the Glasso was for micro-array data, we have successfully applied it to financial time series, demonstrating it's application satisfies fundamental structures (companies with in industries cluster together) and discovering surprising relationships.

References

[1] Friedman, J., Hastie, T., Tibshirani, R.: *Sparse inverse covariance estimation with graphical lasso*, Biostatistics (Dec 12, 2007)

[2] Davidson-Pilon, C., *Least-Squares regression with L1 penalty*, July 2012, http://camdp.com/blogs/least-squares-regression-l1-penalty

[3] T., Tibshirani, R.: *Regression Shrinkage and Selection via the Lasso*, Journal of Royal Statistical Society, Vol 58, Issue 1, (1996), pg. 267-288

[4] Duchi, J., Gould, S., Koller, D.: *Projected Subgradient Methods for Learning Sparse Gaussians*, Conference on Uncertainty in Artificial Intelligence (UAI 2008).

[5] Witten, D., Freidman, J., Simon, N.: *New Insights and Faster Computations for the Graphical Lasso*, Journal of Computational and Graphical Statistics 20(4): 892-900.

[6] Stark, R. *Illustrating partial correlation vs. interaction: what they are and how they differ*, 2010, http://www.integrativestatistics.com/partial.htm